Determining Association Between Categorical Variables

9 WEDNESDAY
2022 FEBRUARY

2022 JANUARY

S	M	T	W	T	F	S
30	31					1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29

WEEK 07
040-325

Bad Avg. Very good Excellent → Ordinal data

Variable = "Food Quality"

Determining and Inferring association

Conditional Probablity

Ex:- Consider B-school which shortlisted 1200 candidates (960 men and 240 women) for its post graduate management program. Out of them, 324 candidates were given offer letters for admission. The data is included here:

	Male	Female	Total
Offers made	288	36	324
Not offered	672	204	876
Total	960	240	1200

→ Contingency Table

- After reviewing the record, a women's forum raised the issue of gender descrimination on the basis that 288 male candidates were offered admission against only 36 female candidates.

Event of randomly selected of male candidates = A

" " " " female " = B

Event of admission offer being made = a A

~~" " randomly selected candidate for admission offer being made = N~~

• A To this, the B-school management replied that it is not a case of descrimination, but was because of the fact that only 240 women candidates appeared for the examination.

A^c be the event that the candidate is not offered admission.

One can see that $Pr(A) + Pr(A^c) = 1$

• Probablity that a randomly observed candidate is a male and is offered the admission. (Joint event)

$$Pr(M \cap A) = 288/1200 = 0.24$$

• Probablity that randomly observed candidate is a male and is not offered the admission.

$$Pr(M \cap A^c) = 672/1200 = 0.56$$

• Similarly,

$$Pr(W \cap A) = 36/1200 = 0.03$$

$$Pr(F \cap A^c) = 204/1200 = 0.17$$

Joint Probablity

In terms of probablities, the previous table can now be written as:

	Male	Female	Total
Offers made	0.24	0.03	0.27
Not offered	0.56	0.17	0.73
Total	0.8	0.2	1.0

□ = Joint probablity

□ = Marginal probablity

Marginal probablity

- What will be $Pr(A|M)$? (Conditional Probability)

The probability of event A happening, given the fact that event M has already taken place

- This conditional probability tell us that we are concerned with admission status of only males!
- We know that out of the 960 male candidates, 288 were offered admission. So probability that a male candidate is offered admission will be $288/960 = 0.3$

- Also observe that:

$$Pr(A|M) = \frac{288}{960} = \frac{288/1200}{960/1200} = \frac{0.24}{0.8} = 0.3 = \frac{Pr(M \cap A)}{Pr(M)}$$

0.24 → Joint probability
0.8 → Marginal Probability

$$\text{Conditional Probability} = \frac{\text{Joint Probability}}{\text{Marginal Probability}}$$

$$Pr(A|W) = \frac{Pr(A \cap W)}{Pr(W)} = \frac{0.03}{0.2} = 0.15$$

Conclusion

- The probability of admission offer given the candidate is a male is 0.3, twice of 0.15 probability of admission offer given by the candidate is a woman.
- Although the use of conditional probability does not, in itself, prove descrimination, there is support for the argument!

Baye's Rule

- Given the new information, we can update our prior beliefs by calculating revised probabilities— this is called the posterior probability.
- Baye's Rule is used to calculate the posterior probability if we have the initial belief (probability) and the additional sample information.

Ex:-

- Suppose that a manufacturer receives same raw material from two different suppliers S_1 and S_2.
- Currently 65% of the raw material comes from S_1 and remaining, 35%, comes from S_2.
- Also, suppose that from the historical data available with the quality assurance department, we know that S_1 has 98% of the supplied raw material of

SUNDAY 13

good quality and S_2 has 95% of the raw material of good equality.

- That is, the probablity of a "Good" quality raw material given that the supplier is S_1 is, $Pr(G|S_1) = 0.98$. And for the second supplier, this probablity is : $Pr(G|S_2) = 0.95$

Q. What is the probablity of the raw material being supplied by S_1 and it being good?

'and' given (Joint Probablity)

Soln:- Joint probablity, ofcourse!

According to the Baye's formular

$$\text{Conditional Probablity} = \frac{\text{Joint Probablity}}{\text{Marginal Probablity}}$$

Joint Prob. = Cond. Prob. × Marg. Prob.

$$Pr(G|S_1) = \frac{Pr(G \cap S_1)}{Pr(S_1)}$$

$$Pr(G \cap S_1) = Pr(G|S_1) \times Pr(S_1)$$
$$= 0.98 \times 0.65 = 0.637$$

$$Pr(G \cap S_2) = Pr(G|S_2) \times Pr(S_2)$$
$$= 0.35 \times 0.95 = 0.3325$$

- Now, knowing all this information so far, suppose the manufacturer inspects the incoming raw material on receipt and finds a bad quality material.

- He wants to know the supplier who needs to be contacted to complain!

- We are interested in the posterior probablity that a particular supplier is guilty of supply bad quality product given that we have bad quality raw material at our doorstep - $Pr(S_1|B)$ or $Pr(S_2|B)$
- This is an application of Baye's theorem — finding posterior probablity given some initial facts and numbers.
- From Baye's formula we know that:

$$Pr(S_1|B) = \frac{Pr(S_1 \cap B)}{Pr(B)}$$

$$= \frac{Pr(S_1) \times Pr(B|S_1)}{Pr(B)}$$

- What is $Pr(B)$?
- That is the probablity of receiving a bad quality raw material.
- Now bad quality raw material can from supplies of S_1 or S_2.

16 WEDNESDAY
2022 FEBRUARY

WEEK 08
047-318

2022 JANUARY

S	M	T	W	T	F	S
30	31					1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29

- That is, the event B can occur with S_1 or with S_2.

$$Pr(B) = Pr(S_1 \cap B) + Pr(S_2 \cap B)$$

- But $Pr(S_1 \cap B) = Pr(S_1) \times Pr(B|S_1)$, and
- $Pr(S_2 \cap B) = Pr(S_2) \times Pr(B|S_2)$

$$Pr(S_1|B) = \frac{Pr(S_1) \times Pr(B|S_1)}{Pr(S_1) \times Pr(B|S_1) + Pr(S_2) \times Pr(B|S_2)}$$

$$= \frac{0.65 \times 0.02}{0.65 \times 0.02 + 0.35 \times 0.05} = 0.426$$

$$Pr(S_2|B) = 0.574$$

- Significance: Find posterior probability using prior information
- Notice that we use $Pr(B|S_1)$ to find $Pr(S_1|B)$

Inferring Association Between Categorical Variables: Chi Squared Test of Independence

MARCH 2022

M	T	W	T	F	S	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

THURSDAY
FEBRUARY 2022
17

WEEK 08
048-317

Example: Brand preferences

- Suppose a survey is conducted in Mumbai and Chennai asking respondents their preferences about three brands. The result is summarized below:

	Preferred Brand			
City	Brand A	Brand B	Brand C	Total
Mumbai	279	73	225	577
Chennai	165	47	191	403
Total	444	120	416	980

- Independent (explanatory) variable is the city.
- Dependent (response) variable is the brand preference

- We know how to summarize the data by calculating the marginal and joint probabilities.
- Two categorical variables are statistically independent if the population condition distribution on one of them is identical to each category of the other.
- In the example, the two conditional distributions are not identical. e.g. Brand A is preferred more in Mumbai than in Chennai.
- Refer to the same example extended to a third city:

	Preferred brand			
City	Brand A	Brand B	Brand C	Total
Mumbai	440 (44%)	140 (14%)	420 (42%)	1000 (100%)
Chennai	44 (44%)	14 (14%)	42 (42%)	100 (100%)
Delhi	110 (44%)	35 (14%)	105 (42%)	250 (100%)

MARCH
APRIL

18 FRIDAY
2022 FEBRUARY

2022 JANUARY

S	M	T	W	T	F	S
30	31					1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29

WEEK 08
049-316

- Conditional distributions is same across the cities. Hence we can conclude that brand preference is independent of the cities.
- However, statistical independence is a ~~summary~~ symmetric property b/w two categorical variable.

- If the conditional distributions within the rows are identical, then so are the distributions within the columns.
- One can verify that the conditional distributions amongst columns equals (74%, 7%, 19%)
- Based on this single sample information, we can draw inferences about the population, as we have been doing?

Chi-Square distribution.

- Null hypothesis -
  H_0: The categorical variables are independent.
- Alternate hypothesis -
  H_1: The categorical variables are not independent.

Let f_o be the observed frequencies (from the sample)
Let f_e be the expected frequencies, if the variables were independent.

The expected frequency for a cell equals the product of row and column totals for that cell, divided by total sample size.

- Calculate the expected frequency:

Brand A (Mumbai) = $\frac{444 \times 577}{980} = 261.4$

Brand A (Chennai) = $\frac{444 \times 403}{980} = 182.6$

Brand B (Mumbai) = $\frac{120 \times 577}{980} = 70.7$

Brand B (Chennai) = $\frac{120 \times 403}{980} = 49.3$

⋮

City	Preferred brand: Brand A (observed frequency)	Brand B	Brand C	Total
Mumbai	279 (261.4)	73 (70.7)	225 (244.9)	577
Chennai	165 (182.6)	47 (49.3)	191 (171.1)	403
Total	444	120	416	980

(Expected frequency)

- Chi-Squared test statistic:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = \frac{(279-261.4)^2}{261.4} + \frac{(73-70.7)^2}{70.7} + \frac{(225-244.9)^2}{244.9}$$
$$+ \frac{(165-182.6)^2}{182.6} + \frac{(47-49.3)^2}{49.3} + \frac{(191-171.1)^2}{171.1} \approx 7.0$$

SUNDAY 20

21 MONDAY
2022 FEBRUARY

WEEK 09
052-313

2022 JANUARY

S	M	T	W	T	F	S
30	31					1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29

- When the Ho is true, expected and observed frequencies tend to be close for each cell, and the test statistic value is relatively small.
- If Ho is false, atleast some cells have a big gap b/w expected and observed frequencies, leading to a large test statistic value.
- The larger the χ^2 value, greater is the evidence against the null hypothesis of independence
- Degrees of freedom for the Chi-Squared distribution is given by the expression : ~~df = (row-1)x(col-1)~~ : $df = (r-1) \times (c-1)$. r and c are the # of rows and columns respectively.

$\chi^2 = 7.0$, $df = 2$, So at $\alpha = 0.05$ (95% confidence), the tabular value of test statistic, $\chi^2 = 5.99$

- So we reject the null hypothesis of independence.
- However, at $\alpha = 0.01$ (99% confidence), the tabular value of test statistic, $\chi^2 = 9.21$, and we can not reject the null hypothesis.

Box Plot	Outlier detection
Line Chart	Trend Analysis
Column Stacked Chart	Proportion Compare parts of a whole
Histogram	To guess distribution
Scatter Plot	Correlation Analysis
Bar Chart	To do comparison
Pie Chart	Proportion

Ho(null hypothesis) :- The population follows the proposed distribution.

Ha(alternative hypothesis) :- The population does not follow the proposed distribution

Q. In the goodness-of-fit test, if the computed test statistic is greater than the tabulated value of the statistic at a given significance level then,- At the specified significance level, reject the null hypothesis and conclude that the data does not come from population with proposed distribution.

Q. A distribution is Left-tailed if...

(a) Negative Skewness
(b) Left Tail implied mean is on left side
(c) Mean < Median < Mode

Poisson distribution – Mean & Variance are same $= \sigma^2$

Exponential – Mean & Median are not same

Standard normal dist. – Mean is 0

Uniform distribution – Mean, Median and Mode are very close to each other

$$\text{Expected frequency} = \frac{\text{Total observed value}}{\text{No. of bins}}$$